ÉCOLE NATIONALE DES CHARTES

Paul, Hector Kervegan

Traitement, exploitation et analyse d'un corpus semi-structuré : le cas des catalogues de vente de manuscrits

Mémoire pour le diplôme de master

Technologies numériques appliquées à l'histoire

2022

Résumé

Introduction

J'ai choisi de structurer mon mémoire autour de plusieurs questions connexes, qui, à différents degrés, se retrouvent tout au long du développement :

En quoi la nature semi-structurée du corpus permet d'en automatiser le traitement ? Comment produire des informations normalisées et exploitables à partir d'un corpus textuel semi-structuré ? En quoi ce traitement et la traduction des documents vers d'autres formats et d'autres médias impacte leur réception ? Quels sont les choix techniques qui influencent cette réception ?

Les deux premières questions, d'orientation plutôt technique, forment la colonne vertébrale pour le mémoire ; elles lient deux aspects centraux : la nature du corpus et la manière dont sa structure permet toute la chaîne de traitement. Par « semi-structuré », j'entends que, à un niveau distant, toutes les entrées de catalogue suivent la même structure ; des séparateurs distinguent les différentes parties, et les informations sont souvent structurées de manière semblable pour chaque manuscrit vendu. Cela permet un traitement de « basse technologie » (*low-tech*) en évitant d'entraîner de lourds modèles de traitement du langage naturel (ce qui aboutirait à des solutions complexes, difficiles à maintenir et à faire évoluer et relativement opaques dans leur fonctionnement). À l'inverse, un corpus semi-structuré peut être traité en déduisant une « structure abstraite », que chaque entrée de catalogue partage. Il est alors possible de mettre en place des solutions techniques plus faciles, pour un résultat de qualité équivalente. Produire des « informations normalisées et exploitables » implique de traiter le corpus en cherchant des réponses à des questions de recherche précises – dans le cadre de mon stage, une question centrale a été de chercher à isoler les facteurs déterminant le prix d'un manuscrit.

Les deux dernières questions, au premier abord plus théoriques, me semblent centrales, notamment à la troisième partie de ce mémoire. Numérisation, traitement informatisé et diffusion sur le web ne sont pas des opérations neutres, mais un ensemble de « traductions » des documents originels. Ces processus comportent une part de choix conscients, qu'il s'agit de mettre en avant. Par exemple, on considère que la majorité des documents vendus ont pour titre l'auteur.ice du document. Cette personne n'est cependant pas toujours mentionnée, et des documents peuvent être nommés d'après un lieu, un évènement ou un thème (la révolution française, par exemple). Ces « traductions » des catalogues sont relativement discrètes tout au long de la chaîne de traitement (où le

format dominant est la `TEI`, qui garde une relation d'équivalence avec le texte). C'est lors du passage au site web que ce processus de traduction devient plus évident, et, potentiellement, plus problématique. On y abandonne la référence au document originel (les catalogues numérisés ne sont pas accessibles en ligne par un.e utilisateur.ice), le catalogue n'est plus la manière privilégiée d'accéder aux items vendus... De plus, la construction d'un site web implique la conception d'une interface et, dans notre cas, la production d'une série de visualisations intégrées au site. Le passage au site web remet aussi en cause la hiérarchie habituelle entre ingénierie et recherche : la conception d'un site ne répond pas à une question scientifique, mais elle soulève ses propres questions. Loin d'être anodines, ces problématiques de design déterminent la construction et la réception des savoirs. Il est donc important, je pense, de problématiser ces questions de visualisation et de design.

Première partie

Du document numérisé au XML-TEI : nature du corpus, structure des documents et méthode de production des données

Chapitre 1

Le marché des manuscrits autographes au prisme des catalogues de vente

Ce chapitre présente l'objet d'étude du projet *MSS* : étudier le marché des manuscrits autographes du XIX$^{\text{ème}}$ s.. parisien à partir de ses catalogues de vente et étudier la construction du canon littéraire au prisme du marché du manuscrit.

1.1 Pourquoi étudier le marché des manuscrits autographes ?

Cette section porte sur l'intérêt scientifique des objets d'étude du projet (marché des manuscrits et étude de la construction du canon).

1.2 La structure du corpus : périodisation, producteurs des documents et classification

Ici est faite une présentation des documents traités dans le cadre du projet MSS. La présentation est à deux niveaux : au niveau du corpus et des catalogues. Ce chapitre s'appuie sur les mémoires effectués par d'ancien.ne.s stagiaires de *Katabase*, qui ont déjà beaucoup analysé la nature et les enjeux du corpus [1].

1. Lucie Rondeau du Noyer, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la revue des autographes de la librairie Charavay*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : `https://github.com/lairaines/M2TNAH` (visité le 13/06/2022) ; Caroline Corbières, *Du catalogue au fichier TEI. Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2020, URL : `https://github.com/carolinecorbieres/Memoire_TNAH`

1.2.1 Le corpus de catalogues de vente de manuscrits

Ici est présenté le corpus : nature, quantité de documents (et d'entrées individuelles), dates, différentes classifications qui peuvent être faites (revues, catalogues de ventes aux enchères, catalogues à prix fixes ou d'enchères...).

1.2.2 Structure des catalogues

Ici sera présentée la structure des catalogues ; la structure de chaque page ne sera détaillée qu'à la partie suivante.

(visité le 13/06/2022) ; Juliette Janès, *Du catalogue papier au numérique. Une chaîne de traitement ouverte pour l'extraction d'information issue de documents structurés*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2021, URL : `https://github.com/Juliettejns/Memoire_TNAH` (visité le 13/06/2022).

Chapitre 2

Production des données : de l'OCR à la TEI

Cette partie s'attache autant à présenter le processus d'océrisation (qui est déjà bien établi et ne constitue pas le cœur de mon stage) que la structure des documents. Alors que le chapitre précédent s'intéresse aux catalogues dans leur ensemble, ici, on étudie le corpus au niveau de la page et de l'entrée individuelle. En effet, l'océrisation repose sur la segmentation, et donc sur l'établissement d'une structure « abstraite » d'une page (c'est-à-dire, d'un découpage de la page en zones).

2.1 Extraire le texte des imprimés

2.1.1 Comprendre la structure du document pour préparer l'édition numérique

Appréhender la structure de la page à l'aide de SegmOnto

La structure des catalogues est présentée au niveau de la page. L'ontologie SegmOnto [1] est utilisée, autant pour appréhender la structure de la page que pour exprimer cette structure de façon standardisée.

Description des entrées de catalogue : préparer l'édition TEI

Ici, la structure des catalogues est présentée au niveau de l'entrée, c'est à dire du lot mis en vente. C'est à autour de la structure des entrées qu'est construite l'édition XML-TEI. On s'intéresse à la structure des entrées individuelles à deux niveaux :

1. Kelly Christensen, Simon Gabay, Ariane Pinche et Jean-Baptiste Camps, « SegmOnto – A Controlled Vocabulary to Describe Historical Textual Sources », dans *Documents anciens et reconnaissance automatique des écritures manuscrites*, Paris : École nationale des Chartes, 2022.

— Au niveau intellectuel : quelles sont les différentes parties d'une entrée (titre, description du manuscrit, prix...).
— Au niveau « textuel » : quels sont les séparateurs, c'est à dire les éléments dans le texte qui permettent de séparer les pages de catalogue en entrées et les entrées en sous-éléments) correspondant à la structure intellectuelle décrite ci-dessus.

2.2 L'encodage des manuscrits en XML-TEI

2.2.1 Encoder les catalogues en TEI

Ici est présentée la représentation XML-TEI des catalogues de vente.

2.2.2 L'encodage en TEI : un processus sélectif qui réduit les significations du texte

Après une étape d'océrisation via eScriptorium, le texte extrait des PDF peut être exporté soit en texte brut, soit en XML Page ou Alto. Ces formats s'attachent à garder une relation entre le XML et le document numérisé (les zones de texte sont indiquées, chaque ligne est dans une balise...). Cependant, l'unité intellectuelle centrale à la suite du projet, ce n'est pas la page numérisée, mais l'entrée de catalogue. Un format plus complexe que le XML d'eScriptorium est donc nécessaire. Assez logiquement, la suite du projet s'appuie sur une traduction des catalogues en TEI. On s'intéresse autant à la structure des documents XML (quelles balises sont utilisées...) qu'à l'intérêt scientifique d'une édition numérique (balisage sémantique, possibilité de normaliser les informations grâce à des attributs).

L'édition numérique en XML-TEI des catalogues implique une certaine perte d'informations : l'intégralité des significations contenues dans les catalogues imprimés ne peut être traduite en TEI (la police, ou la qualité du papier, peuvent être documentés mais ne peuvent pas être reproduites). Ce genre de perte d'information a lieu, à différents degrés, dans la plupart des éditions TEI : ce format n'est pas un substitut des documents originels. Dans le projet *MSS / Katabase*, d'autres informations sont perdues : l'édition numérique n'est pas censée être exhaustive des catalogues. La TEI n'est pas utilisée comme un format de conservation, mais comme un format de traitement qui sera enrichi dans les différentes étapes. Afin de mesurer ce qui est conservé et ce qui est perdu du document originel, l'édition TEI sera analysée à la lumière de la « roue du texte » du philologue Patrick Sahle [2] qui modélise les significations plurielles d'un texte.

2. Patrick Sahle, « Digital modelling. Modelling the digital edition », dans *Medieval and modern manuscript studies in the digital age*, London/Cambridge, 2016, URL : https://dixit.uni-koeln.de/wp-content/uploads/2015/04/Camp1-Patrick_Sahle_-_Digital_Modelling.pdf.

Deuxième partie

Normalisation, enrichissements et extraction d'informations : une chaîne de traitement pour des données semi-structurées

Chapitre 3

Faire sens d'un corpus complexe : homogénéisation des données et extraction d'informations

3.1 Homogénéiser et normaliser un corpus complexe

Cette section s'intéresse à la manière dont les fichiers `TEI` sont traités afin de pouvoir ensuite en extraire des informations. C'est directement grâce la structure des entrées (et grâce à la nature « semi-structurée » des catalogues) qu'est possible le traitement automatisé des documents.

3.1.1 Pourquoi chercher à normaliser le corpus ?

La question mérite d'être posée : si il faut travailler le corpus pour pouvoir en extraire des informations, et pour pouvoir donc en faire sens, ce processus peut également impacter la nature du texte et ses significations. Tout comme l'édition `TEI` originelle est un processus sélectif, le traitement des documents encodés est lui un processus sélectif : certaines informations contenues par l'encodage sont privilégiées aux dépens d'autres. Il s'agit ici d'expliciter ces choix (travailler sur les prix, les formats et dimensions des manuscrits plutôt que sur leur sujet) et de les justifier. Cette section s'attache donc à rappeler les questions de recherche qui sous-tendent la normalisation des documents (ajouter plus de structure au document `TEI` pour l'exploiter plus facilement, uniformiser la notation des tailles et des dimensions des documents...).

3.1.2 Comment normaliser le corpus tout en préservant sa valeur documentaire ?

Ici, on s'intéresse à la manière dont la TEI est mise à profit pour enrichir le corpus tout en conservant le contenu textuel des catalogues. Les différentes étapes de normalisation sont également rappelées (ce travail n'étant pas au cœur de mon stage, il s'agira plutôt d'un rappel que d'une présentation technique détaillée).

3.2 Faire sens du corpus : extraction d'informations et fouille de texte

Ici sont décrits le processus et les objectifs de l'extraction d'informations à partir des fichiers TEI. C'est à partir de cette opération d'extraction que sont construites les visualisations, qui permettent une approche graphique du corpus et une meilleure compréhension de celui-ci.

3.2.1 Extraire des informations au niveau des catalogues

Des données sont extraites pour chaque entrée de catalogue (un travail largement effectué par A. Bartz, que j'ai légèrement mis à jour) : prix (dans la monnaie de l'époque et en francs constants), date de vente normalisée, nom de l'auteur.ice et description du manuscrit... L'extraction d'informations pour les entrées individuelles permet surtout de faire une réconciliation des manuscrits vendus (c'est à dire, de retrouver les items vendus plusieurs fois).

3.2.2 Extraire des informations au niveau des entrées

Un second processus d'extraction produit des données pour chaque catalogue de vente : titre du catalogue, date de vente, nombre d'items vendus, prix minimum, inférieur et maximum, prix moyen et médian, variance... Ce processus permet de produire des données statistiques sur l'évolution du cours et du volume du marché des manuscrits (nombre d'items en vente, évolution des prix).

3.2.3 Vers une approche économique du corpus : la conversion automatique des prix en francs constants

En même temps que ces processus d'extraction d'informations, un script de conversion des monnaies (françaises et étrangères) en francs constants 1900 a été élaboré. En annulant l'effet de l'inflation, les francs constant permettent d'étudier l'évolution réelle des prix.

Chapitre 4

Vers une étude des facteurs déterminant le prix des documents : alignement des entrées du catalogue avec Wikidata et exploitation de données normalisées

Ce chapitre est construit autour d'une question de recherche : comment produire des informations exploitables pour une étude économétrique à partir d'un corpus textuel semi-structuré ? Un des objectifs du projet est de faire l'étude des facteurs déterminant le prix d'un manuscrit. Pour faire cette étude, il faut obtenir, pour chaque entrée du catalogue, un certain nombre d'informations normalisées. Le travail d'extraction de données présentes dans les catalogues a déjà été fait par de précédent.e.s stagiaires. Ces données sont principalement quantitatives : prix des manuscrits, dimensions et nombre de pages, date de création. Il est nécessaire de compléter les informations par des données qualitatives et d'enrichir les données disponibles avec des sources extérieures. Pour ce faire, il a été choisi d'aligner le nom des auteur.ice.s des manuscrits avec des identifiants Wikidata ; dès lors que l'on a un identifiant Wikidata, il est possible de récupérer automatiquement des informations sur les personnes via `SPARQL`. Le choix de travailler uniquement sur les noms, et non sur la description des documents, a deux motivations :

— Les noms de personnes (et la manière dont elles sont décrites) constituent la partie la plus normalisée des documents. La description des manuscrits est plutôt en « texte libre ». Dans la continuité avec le reste du projet, nous sommes resté dans une approche « basse technologie », qui consiste à s'appuyer majoritairement sur des solutions techniquement simples. C'est pourquoi nous avons préférer traiter

les noms avec des tables de correspondance [1] et des *expressions régulières*, plutôt que de faire du TAL sur la description des documents.

— Toutes les informations « simples » (données quantitatives facilement normalisables : dates etc.) ont déjà été extraites des descriptions des manuscrits.

Ce travail d'enrichissement a été fait en deux temps.

La première étape, et la plus difficile, est l'alignement avec Wikidata. Cela demande d'extraire un ensemble d'informations à partir du nom de la personne et de la description de celle-ci. Parmi les informations extraites : nom, prénom, titre de noblesse, occupation, dates de vie et de mort. À partir de ces informations, stockées dans un dictionnaire, un algorithme construit successivement différentes chaînes de caractères à rechercher sur l'API de Wikidata. L'objectif est que le premier résultat recherché sur Wikidata soit correct. Sur un jeu de test, le score F1 [2] obtenu est de 68%. Un relecture « manuelle » des résultats est donc nécessaire.

La deuxième étape, nettement plus simple, consiste à lancer des requêtes Wikidata sur les identifiants récupérés afin de récupérer des informations sur les auteur.ice.s des manuscrits (cette partie du travail est encore en cours) pour enrichir nos données.

Une fois ce travail effectué, l'enrichissement des données à proprement parler est possible : les fichiers `TEI` sont mis à jour pour ajouter les identifiants Wikidata. Ainsi, il est possible de faire le lien entre les entrées de catalogues dans des fichiers `XML` et les données issues de requêtes `SPARQL`, stockées dans un `JSON`.

4.1 Questions introductives : pourquoi et comment s'aligner avec Wikidata ?

Cette section, introductive, répond à des questions évidentes mais essentielles : elles permettent de mettre au clair l'intérêt et les (multiples) difficultés dans l'alignement avec Wikidata.

4.1.1 Pourquoi s'aligner avec des identifiants Wikidata ?

Nos données sont déjà complètes, une pipeline entière existe déjà. Cependant, il peut être difficile de déterminer ce qui fait le prix d'un manuscrit. On aborde les manuscrits avec nos propres catégories intellectuelles du XXI$^{\text{ème}}$ s., et notre connaissance de l'histoire

1. C'est à dire, des tables qui permettent de normaliser la manière dont les informations figurent dans les catalogues, et donc de remplacer des termes « vernaculaires » par leurs équivalents utilisés par Wikidata

2. Le score F1, ou *F-score*, est la moyenne harmonique de la précision (vrais positifs par rapport au total de résultats obtenus) et du rappel (nombre de résultats positifs par rapport au total de résultats positifs). (*Précision et rappel*, Wikipedia. L'encyclopédie libre, 2022, URL : `https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel` [visité le 13/06/2022])

de l'époque. Il n'est pas non plus possible de reconstruire d'une manière exacte le regard qu'un public du XIX^{ème} s. aurait sur ces manuscrits – ce qui permettait de revenir à une perception antérieure de la valeur. Il faut donc chercher à contourner ces biais en produisant des données aussi objectives que possibles. Ainsi, un maximum de variales sont à notre disposition pour calculer des régressions linéaires (qui permettent de prédire l'impact d'une variable sur l'évolution des prix).

4.1.2 Quelles données rechercher via `SPARQL` ?

Dans cette sous-section, nous détaillons les données récupérées via `SPARQL` ainsi que les choix scientifiques qui sous-tendent nos décisions.

4.1.3 Comment traduire des descriptions textuelles datant du XIX^{ème} s. en chaînes de caractères qui puissent retourner un résultat sur Wikidata ?

Ce problème est autant linguistique de technique. Une personne ou une chose est nommée ou décrite d'une certaine manière dans un catalogue de vente ancien. Il n'y a aucune garantie que cette caractérisation corresponde à ce qui est disponible sur Wikidata : l'orthographe des noms évoluent, tout comme la manière de nommer certains métiers. À ces évolutions graphiques s'ajoutent des évolutions intellectuelles : les titres de noblesse sont un marqueur plus important au XIX^{ème} s. français que dans un XXI^{ème} s. mondialisé. Une personne n'est que rarement décrite par son titre dans Wikidata.

4.1.4 Comment négocier avec le moteur de recherche de Wikidata ?

Si les catalogues de vente fonctionnent avec leurs propres catégories mentales, le même peut être dit de Wikidata : certains types de données sont plus souvent référencées que d'autres et Wikidata utilise un vocabulaire qui lui est propre. Par expérience, le moteur de recherche de Wikidata est assez « rigide » : contrairement à un moteur généraliste, il n'admet pas d'orthographes alternatives, par exemple. Le traitement des données textuelles et tout le processus de normalisation des données dépendent de ces faits : il faut trouver quelles informations sont référencées par Wikidata et comment elles le sont. L'algorithme doit donc s'adapter au moteur de recherche.

4.1.5 Une approche prédictive

Étant donnée la quantité d'incertitude présentée ci-dessus, l'approche suivie dans l'alignement avec Wikidata est prédictive. L'objectif de l'algorithme de recherches en

plein texte sur l'API Wikidata n'est donc pas de trouver la « bonne » réponse. Il est de construire une chaîne de caractère dont on prédit qu'elle apportera un résultat pertinent. De la même manière, la phase de préparation des données est un processus qui sélectionne et normalise certaines informations dont on suppose qu'elles seront pertinentes dans l'obtention des bons résultats. De la même manière, le premier rôle des tests est de quantifier les prédictions. Ils répondent à la question : étant donné les résultats obtenus lors des tests, quelle est la probabilité que la prochaine chaîne de caractères recherchée retourne un résultat pertinent ? Cette approche prédictive implique nécessairement un degré d'incertitude, et donc le développement d'algorithmes flexibles qui cherchent à minimiser le bruit.

4.1.6 Présentation générale de l'algorithme

À l'aide d'un schéma, on présente l'intégralité de la *pipeline* pour cette étape : préparation des données ; lancement des recherches en plein texte sur l'API Wikidata ; lancement de requêtes `SPARQL` à partir des résultats obtenus et structuration des résultats.

4.2 Préparer et structurer les données

Avant de chercher à récupérer un identifiant Wikidata via l'API, un algorithme se charge de traduire et de structurer les données : à partir d'un nom et de son éventuelle description, un dictionnaire qui contient les informations de manière structurée est construit. À partir de ce dictionnaire, un algorithme contenant différentes requêtes est lancé pour récupérer les identifiants Wikidata.

4.2.1 Présentation générale

Ici, on présente la pipeline de l'algorithme (à l'aide d'un schéma), les données fournies en entrée et le résultat produit en sortie. Les sections suivantes détaillent quelques points d'intérêts.

4.2.2 Identifier le type de nom

Les éléments `tei:name` contiennent le nom qui est donné à un document. Si c'est souvent un nom de personne, ce n'est pas toujours le cas (il y a aussi des noms de lieux, d'évènements), et il y a plusieurs types de noms de personnes : un nom peut être écrit en suivant différentes structures, ce qui appelle à différents types de traitements.

4.2.3 Reconstruire un prénom complet à partir de son abréviation

Souvent, le prénom d'une personne est écrit en abrégé. Partant de ce constat, un algorithme a été construit pour :

— Repérer lorsqu'un prénom est abrégé, en prenant en compte différents types d'abréviations (nom simple ou composé, nom entièrement ou partiellement abrégé) et des possibles fautes dans les catalogues (un point est oublié à la fin d'une abréviation, par exemple).

— Reconstruire un prénom complet à partir de son abréviation, ce qui passe pas un algorithme qui cherche à reconstruire le nom en plusieurs étapes pour obtenir le nom le plus complet possible avec un mimimum d'erreurs.

Ici, l'approche est totalement prédictive : il est impossible d'être certain d'obtenir le bon nom complet à partir de son abréviation ; on peut uniquement prédire que le prénom reconstruit sera conforme au vrai prénom (tel qu'il est écrit sur Wikidata) et chercher à maximiser cette certitude.

4.2.4 Extraire des informations normalisées à partir d'un nom et de sa description

Cette sous-section détaille l'utilisation de tables de conversion pour traduire et normaliser certaines donnes importantes (dates de vie et mort, titres de noblesse et fonctions).

4.3 Extraire des identifiants Wikidata

Une fois un dictionnaire de données normalisées produites, un algorithme lance des recherches en plein texte sur l'API de Wikidata afin de récupérer des identifiants. L'algorithme lance plusieurs requêtes successivement. L'objectif est de récupérer un identifiant en lançant le moins de requêtes, avec le plus de certitude possible.

4.3.1 Présentation générale

Ici est présenté le fonctionnement général de l'algorithme, qui se comporte différemment en fonction du type de données qu'il a à traiter (personne noble ou non, prénom reconstruit ou non...)

4.3.2 Gérer la montée en charge : optimisation et réduction du temps d'exécution

Le script est assez compliqué, repose sur une API et traite un grand nombre de données (plus de 82000 entrées). Il prend donc plus d'une dizaine d'heures à s'exécuter et demande des ressources élevées (la première version du script ne fonctionnait plus sur mon ordinateur après avoir traité 5% du jeu de données). L'optimisation nécessaire de l'algorithme est décrite dans cette sous-section.

4.3.3 Évaluation du script : tests, performance et qualité des données extraites de Wikidata

Des tests ont été réalisés pour :

— isoler l'impact de chaque paramètre (élément du dictionnaire) dans l'obtention des bons résultats
— évaluer la qualité de l'algorithme final
— mesurer la performance de celui-ci.

Ces tests, et leurs résultats, sont présentés ici.

4.4 Après l'alignement, l'enrichissement : utiliser `SPARQL` pour produire des données structurées

4.4.1 Produire des données exploitables via `SPARQL`

La récupération des identifiants Wikidata est la partie la plus complexe dans l'utilisation de Wikidata pour enrichir des données. Après une présentation des informations requêtées via `SPARQL`, le processus d'extraction d'informations et de stockage dans un `JSON` est détaillé.

Lier la `TEI` aux données nouvellement produites

Cette courte section détaille la mise à jour des fichiers `TEI` avec les identifiants Wikidata, ce qui permet de faire le lien entre les entrées de catalogues et les données issues de Wikidata.

4.5 Des données à la monnaie : premiers résultats de l'étude

Sous réserve que l'étude des régressions linéaires ait été fait à temps (ce qui n'est pas garanti), j'aimerais ici présentés les premiers résultats sur les facteurs de l'évolution des prix.

Troisième partie

Après la TEI : l'application web *Katabase*, interface de diffusion des données

Chapitre 5

Design d'interface dans un projet d'humanités numériques : l'application web *Katabase*

Ce chapitre s'intéresse aux relations entre *web design*, données textuelles et humanités numériques, à partir de l'exemple du site web développé pour le projet *Katabase*.

5.1 Le design d'interfaces : une reconfiguration des méthodes de recherche et une transformation du corpus

Cette section s'intéresse aux nouveautés apportées par le design d'interfaces dans les humanités numériques. On s'intéresse à la manière dont le design d'interfaces (et le design de façon générale) transforme les méthodes de recherche « habituelles », mais aussi une transformation du rapport aux documents.

5.1.1 Le design comme inversion des méthodes

Avec les humanités numériques, les questions de design et de structuration deviennent centrales, depuis la conception de schémas `TEI` (qui demandent de mettre en forme un document pré-existant) et d'ontologies jusqu'au développement d'interfaces et de sites web. Parmi ces questions « formelles », le design d'interfaces occupe cependant une place particulière. En effet, dans la plupart des aspects des humanités numériques, le rapport entre questions techniques et scientifiques est clairement établi ; la question scientifique préexiste, et la technique sert surtout à répondre à cette question (comme cela a été le cas jusqu'à dans la « pipeline » jusqu'ici). Cette hiérarchie entre théorie et

pratique reste somme toute assez traditionnelle et correspond aux méthodes scientifiques établies.

Avec le design d'interfaces, cependant, ce rapport établi s'inverse. En effet, le design ne cherche pas à répondre à une question. Tout au plus, il répond à un cahier des charges (le design doit, à minima, permettre de diffuser des données de façon lisible par des êtres humains). C'est avec la pratique du design que naissent les problématiques, parmi lesquelles :

— Comment organiser les différentes parties d'une page pour que celle ci soit lisible ?
— Comment organiser la relation entre les pages pour qu'un site web soit facilement navigable ?
— De quelle manière l'apparence d'un site détermine la réception des contenus ?
— En quoi le design d'un site web construit ou bouscule des habitudes et des formes d'utilisation chez ses utilisateur.ice ?

Toutes les questions posées par le design n'attendent pas nécessairement de réponse. Cependant, force est de constater que ce domaine appelle à une nouvelle approche pour des chercheur.euse.s et ingénieur.e.s issu.e.s des humanités ; ces questions visuelles amènent à une approche semblable à celle de la recherche-création et demandent de développer un nouveau rapport à la technique.

5.1.2 Interface et document

En plus de perturber nos méthodes, la conception d'interfaces influence la perception des documents. Dans le cas du projet *Katabase*, le site web opère une médiation, il implique de une « scénographie » autour des catalogues de vente. Ceux-ci et les manuscrits qui y ont décrits sont intégrés à des pages, inclus dans un parcours, accessibles depuis différents points d'entrée. En plus de cette scénographie, les catalogues sont littéralement traduits, depuis la `TEI` vers le format `HTML`, ce qui implique une perte d'information (les métadonnées du `teiHeader`). Enfin, le site internet marque avant tout un éloignement intellectuel avec les documents : le catalogue n'y est plus l'unité intellectuelle dominante, alors qu'il restait l'un des critères structurants des fichiers `TEI` (un fichier représentant un catalogue). Sur le site web, on peut accéder directement aux éléments vendus, sans avoir à passer par les catalogues. Dans le contexte d'un projet issu de la littérature, toutes ces opérations ne sont pas neutres et méritent d'être explicitées. La méthode de traduction de `XML-TEI` vers `HTML` peut également être présentée ici.

5.2 La conception d'interface, un problème pour les humanités numériques ?

Cette section s'intéresse aux rôle des interfaces en humanités numériques.

5.2.1 Pour une approche pragmatique du design d'interfaces dans un contexte d'humanités numériques

Le design graphique demande des compétences spécifiques qui ne font pas directement partie des cursus d'humanités numériques. Il ne sert pas seulement à faire des sites qui soient « beaux », il joue un rôle essentiel en encadrant la réception des contenus présentés. Cependant, les approches plus « élaborées » de design d'interfaces demandent des financements et des techniques qui sont souvent hors de portée d'un projet universitaire. Des approches plus « critiques » du design ont également été développées dans les humanités numériques [1]. Ces approches ont tendance à être difficiles à mettre en œuvre ; leur portée critique peut aller à l'encontre de l'utilité des interfaces, en faisant de l'interface l'objet principal d'intérêt, aux dépends des contenus présentés.

À l'opposé de ces approches, ce qui est défendu dans le cadre du projet *MSS / Katabase* est une approche à la fois informée et pragmatique du *web design*. Informée, car être conscient des enjeux du design permet un meilleur positionnement en tant qu'ingénieur.e, et donc une présentation des contenus plus intéressante. Pragmatique, parce que les solutions qui sont présentées sont des solutions techniquement réalisables dans le cadre d'un projet universitaire. C'est ici qu'est présentée la charte graphique développée pour l'application web *Katabase*.

5.2.2 Rejeter les interfaces ?

Après avoir parlé de l'intérêt des interfaces et présenté l'approche suivie au sein du projet *MSS / Katabase*, cette partie s'attache à développer une critique des interfaces. À partir d'une approche historique des interfaces graphiques, des contextes dans lesquelles elles se sont développées, nous revenons sur les concepts centraux à leur développement que sont la notion d'utilisateur et de design d'expérience. Il ne s'agit pas de remettre en cause l'utilisation d'interfaces, mais de défendre une approche critique et consciente de l'impact que la standardisation des « expériences utilisateur » sur internet peuvent avoir sur la diffusion des connaissances.

1. Johanna Drucker, *Visualisation : l'interprétation modélisante*, Paris, 2020 (Esthétique des données, 03).

Chapitre 6

Donner à voir un corpus textuel

Ce chapitre s'intéresse aux visualisations développées pour l'application web *Katabase*.

6.1 Visualisation, design et sciences : des relations complexes

Ici, on s'intéresse à la place qu'occupe la visualisation de données dans la recherche scientifique. Le rapport entre les sciences et la visualisation est loin d'être simple et unidirectionnel : cette dernière n'est pas juste un outil, une méthode utilisée dans la recherche scientifique pour des raisons pratiques. Il est plus intéressant de penser la visualisation (et donc le design) et les sciences comme des domaines en interaction, qui s'influencent mutuellement. De la même manière que l'écriture implique des manières de penser particulières [1] en donnant au discours une existence spatiale (le texte est répandu sur une page et des renvois peuvent être fait d'un endroit de la page à un autre), la visualisation implique ses propres manières de penser et influence donc la recherche. Dans notre cas par exemple, produire des visualisations implique de s'intéresser à des informations quantifiables ; cela encourage donc une approche statistique du corpus. À l'inverse, la recherche scientifique ne fait pas qu'« utiliser » le design. Certaines pratiques sont favorisées et deviennent force d'autorité dans des disciplines scientifiques. Se créent alors des « cultures visuelles » [2] propres à ces disciplines, avec leurs traditions et motifs.

1. Anthony Masure, *Design et humanités numériques*, Paris, 2017 (Esthétique des données, 01), p. 111-116.

2. Klaus Hentschel, *Visual cultures in science and technology. A Comparative History*, Oxford, 2014, p. 14.

6.1.1 L'utilisation de supports visuels dans les sciences : une longue histoire

Ici, on retrace une histoire de l'utilisation du visuel dans les sciences (au sens large : sciences « dures » et sciences humaines), à partir (entres-autres) du travail de Anne-Lyse Renon [3] et de K. Hentschel [4].

6.1.2 Une vision objective ? Visualisation et prétention à l'objectivité

Cette partie fait un retour sur la manière dont le visuel et la production de graphiques ont été utilisés comme arguments d'autorité, afin de montrer des faits de façon « objective » [5].

6.1.3 La tendance visuelle des humanités numériques

Pour finir, on fait un bref retour sur la manière dont les humanités numériques « intensifient » la tendance à la visualisation, ou complexifient le rapport entre sciences et visualisation, pour deux raisons. En premier lieu, les humanités numériques viennent avec le développement de nouveaux outils. Ensuite, les humanités numériques marquent un retour à une approche quantitative et graphique dans les sciences humaines – approche qui trouve ses sources, entre-autres, dans l'École des Annales et sa collaboration avec le Laboratoire de Graphique de Jacques Bertin [6], ainsi que dans le structuralisme, où les « structures » trouvent leur meilleure représentation sous forme graphique. Cette tendance visuelle des humanités numériques n'est pas sans poser problème, puisque les visualisations sont développées par des personnes qui n'ont pas nécessairement de formation en graphisme. Une approche pragmatique de la visualisation a tendance à primer (les graphiques servent à prouver quelque chose), plutôt qu'une approche critique (les représentations graphiques sont des interprétations, où les données sont signifiantes, mais où les formes et les méthodes de représentation importent aussi).

---

3. Anne-Lyse Renon, *Design et esthétique dans les pratiques de la science*, Thèse de doctorat, Paris, École des Hautes Études en Sciences Sociales, Institut Marcel Mauss, 2016, URL : `https://www.academia.edu/36754513/Design_et_esth%C3%A9tique_dans_les_pratiques_de_la_science` (visité le 08/06/2022), p. 47-88.

4. K. Hentschel, *Visual cultures in science and technology. A Comparative History…*

5. A.L. Renon, « "Design graphique" et "objectivité", la question des méta-altas », dans *Voir l'architecture. Contribution du design à la construction des savoirs*, dir. Annick Lantenois et Gilles Rouffineau, Paris, Grenoble, 2015, p. 71-81.

6. Olivier Orain, « Le Laboratoire de cartographie dans le contexte de développement des sciences sociales et humaines, des années 1950 aux années 1970 », dans *Design graphique, recherche et patrimoine des sciences sociales. Le Laboratoire de graphique de Jacques Bertin*, Pierrefitte-sur-Seine : Archives nationales, 2021, URL : `https://dai.ly/x85jbir` (visité le 14/07/2022).

6.2 Interpréter le corpus de manuscrits

Ce chapitre s'intéresse à la manière dont le corpus de catalogues de vente de *Katabase* a été traduit en graphiques et à la manière dont ces représentations permettent un nouveau regard sur le corpus.

6.2.1 La visualisation comme objet de connaissance

Ici, on développe une analyse du corpus de catalogues et des manuscrits qui y sont décrits à partir des visualisations produites. Par leur capacité à traduire les informations sous des formes synthétiques [7], les visualisations sont des objets de connaissance qui permettent de comprendre le corpus traité.

6.2.2 La visualisation comme interprétation

Les représentations graphiques ne font pas que montrer des phénomènes. Leur rôle est moins analytique que démonstratif : elles ne révèlent pas une information qui serait cachée dans les données, mais interprètent celles-ci conformément à une problématique de recherche [8]. Représenter un jeu de données, c'est donc le lire, l'interpréter en fonction de certaines questions scientifiques. Ce processus interprétatif est donc partiel (on ne dit pas tout ce qui est dans un jeu de données, mais seulement ce qui est pertinent dans un certain contexte) ; il est aussi influencé par les propriétés graphiques des visualisations. Cette sous-section s'intéresse donc, à partir d'exemples concrets, à la manière dont les propriétés graphiques (choix de formes et de couleurs) ainsi que le pré-traitement des données et d'autres décisions techniques (représentation des prix en francs courants ou constants) influencent la lecture et la perception du corpus.

7. K. Hentschel, *Visual cultures in science and technology. A Comparative History…*, p. 36.
8. J. Drucker, *Visualisation : l'interprétation modélisante…*, p. 78.

Bibliographie

Projet *MSS / Katabase*

Corbières (Caroline), *Du catalogue au fichier TEI. Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2020, URL : `https://github.com/carolinecorbieres/Memoire_TNAH` (visité le 13/06/2022).

Gabay (Simon), Rondeau du Noyer (Lucie) et Khemakhem (Mohamed), « Selling autograph manuscripts in 19th c. Paris : digitising the Revue des Autographes », dans *IX Convegno AIUCD*, Milan, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02388407` (visité le 13/06/2022).

Gabay (Simon), Rondeau du Noyer (Lucie), Gille Levenson (Matthias), Petkovic (Ljudmila) et Bartz (Alexandre), « Quantifying the Unknown : How many manuscripts of the marquise de Sévigné still exist ? », dans *Digital Humanities DH2020*, Ottawa, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02898929/` (visité le 13/06/2022).

Janès (Juliette), *Du catalogue papier au numérique. Une chaîne de traitement ouverte pour l'extraction d'information issue de documents structurés*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2021, URL : `https://github.com/Juliettejns/Memoire_TNAH` (visité le 13/06/2022).

Khemakhem (Mohamed), Romary (Laurent), Gabay (Simon), Bohbot (Hervé), Frontini (Francesca) et Luxardo (Giancarlo), « Automatically Encoding Encyclopedic-like Resources in TEI », dans *The annual TEI Conference and Members Meeting*, Tokyo, 2018, URL : `https://hal.archives-ouvertes.fr/hal-01819505` (visité le 13/06/2022).

— « Information Extraction Workflow for Digitised Entry-based Documents. » Dans *DARIAH Annual event 2020*, Zagreb, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02508549` (visité le 13/06/1997).

Rondeau du Noyer (Lucie), *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la revue des autographes de la librairie Cha-*

ravay, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : https://github.com/lairaines/M2TNAH (visité le 13/06/2022).

Rondeau du Noyer (Lucie), Gabay (Simon), Khemakhem (Mohamed) et Romary (Laurent), « Scaling up Automatic Structuring of Manuscript Sales Catalogues », dans *TEI 2019 : What is text, really ? TEI and beyond*, Graz, 2019, URL : https://hal.archives-ouvertes.fr/hal-02272962 (visité le 13/06/2022).

Édition numérique, traitement automatisé et analyse de texte

Christensen (Kelly), Gabay (Simon), Pinche (Ariane) et Camps (Jean-Baptiste), « SegmOnto – A Controlled Vocabulary to Describe Historical Textual Sources », dans *Documents anciens et reconnaissance automatique des écritures manuscrites*, Paris : École nationale des Chartes, 2022.

Sahle (Patrick), « Digital modelling. Modelling the digital edition », dans *Medieval and modern manuscript studies in the digital age*, London/Cambridge, 2016, URL : https://dixit.uni-koeln.de/wp-content/uploads/2015/04/Camp1-Patrick_Sahle_-_Digital_Modelling.pdf.

— « What is a scholarly digital edition ? », dans *Digital scholarly editing. Theories and practices*, dir. Matthew James Driscoll et Elena Pierazzo, Cambridge, 2016, URL : http://books.openedition.org/obp/3397 (visité le 10/07/2022).

Visualisation et design d'interfaces

Agamben (Giorgio), *Qu'est-ce qu'un dispositif ?*, Stanford, 2006 (Rivages Poche / Petite Bibliothèque).

Albouy (Ségolène), *Médiation des données de la recherche. Élaboration d'une plateforme en ligne pour une base de tables astronomiques anciennes.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : https://github.com/Segolene-Albouy/Memoire-TNAH2019 (visité le 13/06/2022).

Drucker (Johanna), *Visualisation : l'interprétation modélisante*, Paris, 2020 (Esthétique des données, 03).

Hentschel (Klaus), *Visual cultures in science and technology. A Comparative History*, Oxford, 2014.

Masure (Anthony), *Design et humanités numériques*, Paris, 2017 (Esthétique des données, 01).

Orain (Olivier), « Le Laboratoire de cartographie dans le contexte de développement des sciences sociales et humaines, des années 1950 aux années 1970 », dans *Design graphique, recherche et patrimoine des sciences sociales. Le Laboratoire de graphique de Jacques Bertin*, Pierrefitte-sur-Seine : Archives nationales, 2021, URL : `https://dai.ly/x85jbir` (visité le 14/07/2022).

Renon (Anne-Lyse), « "Design graphique" et "objectivité", la question des méta-altas », dans *Voir l'architecture. Contribution du design à la construction des savoirs*, dir. Annick Lantenois et Gilles Rouffineau, Paris, Grenoble, 2015, p. 71-81.

— *Design et esthétique dans les pratiques de la science*, Thèse de doctorat, Paris, École des Hautes Études en Sciences Sociales, Institut Marcel Mauss, 2016, URL : `https://www.academia.edu/36754513/Design_et_esth%C3%A9tique_dans_les_pratiques_de_la_science` (visité le 08/06/2022).

Économétrie et statistiques

Piketty (Thomas), *Les hauts revenus en France au XXe siècle : inégalités et redistributions, 1901-1998*, Paris, 2001.

Précision et rappel, Wikipedia. L'encyclopédie libre, 2022, URL : `https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel` (visité le 13/06/2022).

Table des matières